PCT/IB04/50050

Europäisches Patentamt | European Patent Office | Office européen des brevets

IB04/50050

# Bescheinigung | Certificate | Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

**Patentanmeldung Nr. Patent application No. Demande de brevet n°**

03075226.5



**PRIORITY DOCUMENT**

SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Anmeldung Nr:
Application no.: 03075226.5
Demande no:

Anmeldetag:
Date of filing: 23.01.03
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

Koninklijke Philips Electronics N.V.
Groenewoudseweg 1
5621 BA Eindhoven
PAYS-BAS

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

Capacity bounds and construction for reversible data-hiding

In Anspruch genommene Priorität(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/
Classification internationale des brevets:

G11B20/00

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IT LU MC NL PT SE SI SK TR LI

# Capacity bounds and constructions for reversible data-hiding

Ton Kalker[a] and Frans M.J. Willems[b]

[a]Philips Research, Prof. Holstlaan 4, Eindhoven, The Netherlands
[b]Technical University Eindhoven, Den Dolech 2, Eindhoven, The Netherlands

## ABSTRACT

An undesirable side effect of many watermarking and data-hiding schemes is that the host signal into which auxiliary data is embedded is distorted. Finding an optimal balance between the amount of information embedded and the induced distortion is therefore an active field of research. In recent years, with the rediscovery of Costa's seminal paper *Writing on Dirty Paper*, there has been considerable progress in understanding the fundamental limits of the capacity versus distortion of watermarking and data-hiding schemes. For some applications, however, no distortion resulting from auxiliary data, however small, is allowed. In these cases the use of reversible data-hiding methods provide a way out. A reversible data-hiding scheme is defined as a scheme that allows complete and blind restoration (i.e. without additional signaling) of the original host data. Practical reversible data-hiding schemes have been proposed by Fridrich et al., but little attention has been paid to the theoretical limits. Some first results on the capacity of reversible watermarking schemes have been derived in[1] and.[2] The reversible schemes considered in most previous papers have a highly fragile nature: in those schemes, changing a single bit in the watermarked data would prohibit recovery of both the original host signal as well as the embedded auxiliary data. It is the purpose of this paper to *repair* this situation and to provide some first results on the limits of *robust* reversible data-hiding. Admittedly, the examples provided in this paper are toy examples, but they are indicative of more practical schemes that will be presented in subsequent papers.

**Keywords:** Watermarking, Data-hiding.

## 1. INTRODUCTION

In 1999 it was observed that data-hiding is closely related to the information-theoretical concept of "channels with side-information". E.g. Chen,[3] Chen and Wornell,[4] and Moulin and O'Sullivan[5] realized that (in the Gaussian case) there is a connection between data-hiding and Costa's *writing on dirty paper*.[6] Costa's achievability proof can be seen as a special case of the proof of Gelfand and Pinsker.[7] Heegard and El Gamal[8] studied codes based on Gelfand-Pinsker theory for computer *memories* with defects. Coding theorems for data-hiding situations appeared in Chen[3] (specialized to the Gaussian case), Moulin and O'Sullivan,[5] Barron,[9] and Willems.[12]

In the present paper we will focus on data-hiding schemes that are robust and reversible, i.e. data-hiding schemes with the additional constraint that the original host signal can be restored from the received signal even under channel degradations. Reversible data-hiding schemes are important in cases where no degradation of the original host signal is allowed. This is for example true for medical imagery, military imagery and multimedia archives of valuable original works.

The literature on reversible data-hiding is not very extensive yet and focusses mostly on *fragile* watermarking schemes. A good overview of the history and the state-of-the-art of reversible data-hiding can be found in Fridrich et al.,[11] Referring to Figure 1 in the latter paper,[11] the general idea of current methods is simple. A set $B$ of features of a signal $X$ is derived such that (i) $B$ can be losslessly compressed, and such that (ii) randomization of $B$ has *little* impact. Lossless data-hiding is then achieved by lossless compression of $B$, concatenating the bitstream with auxiliary data and replacing the original set $B$. Most of the results available in literature focus on practical methods and have little information theoretic aspects. A first attempt at deriving theoretical bounds was made by Kalker and Willems.[1] In most of these previous works, no channel degradations were included (allowed), and the reversible watermarking schemes were highly fragile. This puts a severe limitation on the usability of reversible watermarking schemes: only in a context in which the owner has complete control over the watermarked data (e.g. archives) or in the context of authentication do these watermarking schemes have a useful application.

Authors email addresses: ton.kalker@ieee.org, f.m.j.willems@tue.nl

In this paper we want extend the results from Kalker and Willems [4] to non-fragile, i.e. robust, reversible watermarking schemes. There are a number of possibilities for extending fragile reversible watermarking to robust reversible watermarking. Firstly, robustness can refer to robustness of the watermark payload, i.e. the channel degradations do not interfere with payload recovery. Secondly, robustness can refer to the reversibility aspect, i.e. the original host signal can still be recovered after channel degradations. This second option can be further detailed with respect to the degree with which the original can be restored. At one extreme the original is completely recoverable; at the other extreme the original can only be retrieved up to a *distortion* that is compatible with the channel degradations. Thirdly and finally, robustness can refer to *both* payload and reversibility. The first and second option have limited applicability, as one of two the desirable properties of reversible watermarking is lost (payload or reversibility). This paper therefore focusses on the third option, where robustness refers to both to the payload and the reversibility aspect. In this paper we have chosen for a strict interpretation of reversibility, viz. complete restoration even in the context of channel degradations.

The paper is organized as follows. In Section 2 we set the notation. In Section 3 we provide a simple example that achieves robustness through the addition of parity check bits. In Section 4 we state and informally proof a first result for binary sources. Section 5 states the main result and gives an outline of the proof. Applying the main result to the simple example shows that we can do better than the straightforward compression, adding parity check bits and filling the gap with auxiliary data. The coding scheme presented in Section 7 is inspired by the methods introduced in van Dijk and Willems [10] and Kalker and Willems.[1]

## 2. NOTATION

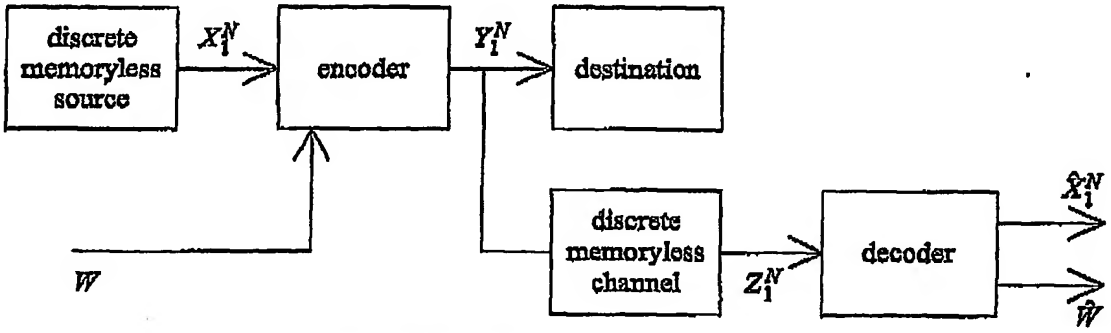


Figure 1. Reversible data-hiding: encoding and decoding.

With reference to Figure 1 we give a formal setup for reversible data-hiding. A source produces the host sequence $x_1^N = x_1 x_2 \cdots x_N$ of symbols from the discrete alphabet $\mathcal{X}$. We assume that the source is memoryless hence

$$\Pr\{X_1^N = x_1^N\} = \prod_{n=1,N} P(x_n), \tag{1}$$

for some probability distribution $\{P(x) : x \in \mathcal{X}\}$. The message source produces the message index $w \in \{1, 2, \cdots, M\}$ with probability $1/M$, independent of $x_1^N$. The encoder (embedder) forms the composite sequence $y_1^N = y_1 y_2 \cdots y_N$ of symbols from the discrete alphabet $\mathcal{Y}$, hence

$$y_1^N = f(x_1^N, w). \tag{2}$$

We now require that the sequences $y_1^N$ must be *close* to $x_1^N$, i.e. the average distortion

$$D_{\text{av}} \triangleq \sum_{x_1^N, w} \frac{1}{M} \Pr\{X_1^N = x_1^N\} d(x_1^N, f(x_1^N, w)) \tag{3}$$

should be *small*. Here

$$d(x_1^N, y_1^N) \triangleq \frac{1}{N} \sum_{n=1,N} D(x_n, y_n) \tag{4}$$

for some specified distortion measure $D(\cdot,\cdot)$. The embedding-rate $R$, in bits per source-symbol, is defined as

$$R \triangleq \frac{1}{N}\log_2(M). \tag{5}$$

The composite sequence is sent through a memoryless *attack* channel with transition probability matrix $Q(\cdot|\cdot)$ to produce a degraded version $z_1^N$ of the watermarked sequence $y_1^N$, i.e.

$$\Pr\{Z_1^N = z_1^N | Y_1^N = y_1^N\} = \prod_{n=1,N} Q(z_n|y_n). \tag{6}$$

The word attack channel is somewhat of a misnomer, as it suggests the presence of an active and intelligent attacker. However, in this paper no such connotation is intended and the word 'attack' is only chosen to reflect common terminology in watermarking literature. From the composite sequence $z_1^N$ the embedded message can be reconstructed reliably, i.e. the decoder produces a message-estimate $\hat{w} = g'(z_1^N)$ such that

$$P_e' \triangleq \Pr\{\hat{W} \neq W\}, \tag{7}$$

is small. Moreover the decoder has to produces an estimate of the host sequence $\hat{x}_1^N = g''(z_1^N)$ such that

$$P_e'' \triangleq \Pr\{\hat{X}_1^N \neq X_1^N\}, \tag{8}$$

is small.

## 3. A FIRST EXAMPLE

Consider the case of a memoryless source with binary alphabet $\mathcal{X} = \{0, 1\}$ with Hamming distance as distortion measure. Let $p_0 = \Pr\{X = 0\}$ and $p_1 = \Pr\{X = 1\}$. Let the attack channel be given as a binary symmetric channel with $0 \to 1$ transition probability equal $d$. In this case it is *theoretically and asymptotically* easy to construct a robust reversible data-hiding scheme with distortion $D_{wv} = 0.5$. Starting with a string $x_1^N$ of length $N$, we first compress the string into a string $y_1^K$, where $K$ is approximately equal to $Nh(p_1)$, where $h(p_1)$ denotes *binary entropy*. To this compressed string $y_1^K$ we add $RN$ auxiliary bits to obtain a sequence $y_1^L$, where $R$ is the rate of the robust reversible data-hiding scheme. To this string $y_1^L$ we add the parity check bits of an appropriately chosen error correcting code $C$ such that the total string has length $N$ (the original length) and such that $dN$ random errors can be corrected. The associated decoding procedure is a simple inversion of the embedding procedure. Firstly, the degraded sequence $z_1^N$ is subjected to error correcting decoding, thereby restoring the sequence $y_1^L$. Secondly, the sequence $y_1^L$ is decompressed until a sequence of length $N$ is obtained. The remaining bits are then automatically obtained as auxiliary message bits. Apart from the initial error correcting step, this decoding procedure is similar to the fragile example as presented in the precursor paper. [1]

It is quite easy to show that for $N$ large, there exists error correcting codes such that the number of parity check bits that has to be added is equal to $Nh(d)$. This leads to the following equation for the robust reversible rate $R_{rev}(p_1, d)$.

$$Nh(p_1) + NR_{rev}(p_1, d) + Nh(d) = N. \tag{9}$$

Solving for $R$, we derive that for $N$ large, and a for binary symmetric attack channel $A$ a rate distortion pair $(R, D)$ can be derived with $R = 1 - h(p_1) - h(d)$ and $D = 0.5$. Comparing with the result from Kalker and Willems [1] we see that robustness requirement creates a loss in rate equal to $h(d)$. The argument also shows that robustness cannot be achieved for attack channels for which $h(d) > 1 - h(p_1)$.

This construction can be slightly generalized by time-sharing, i.e. by performing the construction above on only a fraction $\alpha$ of the symbols in $x_1^N$. The resulting distortion and information rate are then given $D_{wv} = \alpha/2$ and $R = \alpha(1 - h(p_1) - h(d))$. It is to be noted that in this time-sharing construction the parity check bits for *total* string are encoded in the fraction that is being compressed. In summary, asymptotically we can achieve a rate-distortion line

$$\rho_{rev}(p_1, d, \Delta) = 2\Delta(1 - h(p_1) - h(d)), \tag{10}$$

whenever the righthand side of the equation above is positive. Apart from the inclusion of parity check bits, the above method of robust reversible data-hiding is essentially the same method as being proposed by Fridrich et al. [11] and Kalker and Willems.[1] In the latter paper it was shown that for an attack channel equal to the identity the time-sharing was not optimal. The obvious question that we address in this paper is the then following: can we do better than time-sharing in the context of non-trivial attack channels? In the sections below we prove that in general the result as given in Equation 10 is not optimal.

## 4. PRELIMINARY RESULT

In this section we state a preliminary achievability result for binary sources and memoryless channels as in the previous section. A formal proof will not be given as it follows directly from the more general result in the following section. However, we will sketch an argument that cover the basic ingredients of a more formal proof.

**Theorem 1.** *For a binary source $P$ and a binary symmetric attack channel $Q$ with transition probability $d$ as in the previous Section 3, an achievable rate $R$ for a given distortion $\Delta$ is given by*

$$R(\Delta) = \max H(Y) - H(X) - h(d), \tag{11}$$

*where the maximum is over all test-channels $P(y|x)$ such that the average distortion is less than $\Delta$, $\mathrm{E}_{xy}[D(x,y)] \leq \Delta$.*

*Proof.* We sketch a proof of the fact that $R(\Delta)$ as given by the right hand side of Eq. 11 is achievable. For a graphical representation see Figure 2.

Consider a robust reversible watermarking scheme satisfying the conditions of the theorem, and a certain a test channel $P(y|x)$. Assume that this test channel $P$ satisfies the distortion constraint as stated in the theorem. In order to analyze the performance of the coding scheme it is well known that it is sufficient to only consider the set of most likely sequences, the so called *typical* sequences. For large sequence lengths $N$ there are in the order of $2^{NH(X)}$ and $2^{NH(Y)}$ typical sequences $x_1^N$ and $y_1^N$, respectively. Typical sequences $y_1^N$ are only observed through the attack channel $Q$, which introduces a sphere of uncertainty of size $2^{Nh(d)}$ in the space of typical $y$-sequences. Maximum rate reversible embedding is achieved by having the space of typical $y$-sequences $\mathcal{A}(Y)$ as a fibre space over the set of typical $x$-sequences $\mathcal{A}(X)$, where each fibre is partitioned into spheres of size $2^{Nh(d)}$. Each sphere is labeled with a message index $m$. An observer of a $z$-sequence $z_1^N$, traces back to a point in $\mathcal{A}(Y)$ and derives the embedded message by reading the label of the sphere; the original host is reconstructed by projecting on the space $\mathcal{A}(X)$. The rate of the reversible watermarking scheme is now easily computed by counting the number of uncertainty spheres along a fibre. As the number of points on a fibre are given by $\#\mathcal{A}(Y)/\#\mathcal{A}(X) = 2^{N(H(Y)-H(X)}$, the number of spheres is given by $2^{N(H(Y)-H(X)-h(d)}$ whenever the exponent is positive. It follows that the maximum achievable rate is given by $H(Y) - H(X) - h(d)$. □

This result should be compared with the result of[1] on the achievable rates for fragile reversible watermarking schemes for binary sequences. Both results only differ in the term $h(d)$, a constant for a given attack channel. In particular, this preliminary result seems to imply that for both the fragile and the robust (non-fragile) case the maximum rate is achieved for the same test channel. However it certainly shows that the time-sharing construction, as in the fragile case, is not optimal. Recalling the result from,[1] the theorem above implies that for low distortions and small $d$ the optimal test channel is asymmetric, viz. a Z-channel.

## 5. MAIN RESULT

This section states and proves the main result of this paper. The theorem below gives a general result for achievable rates for reversible data-hiding in the presence of attack channels. For the trivial attack channel it reduces to the result of. [1]

A distortion-rate pair $(\Delta, \rho)$ is called achievable if, for all $\epsilon > 0$ there exist, for all large enough $N$, encoders and decoders such that their average distortion, embedding rate, and error probabilities satisfy

$$\begin{aligned} D_{av} &\leq \Delta + \epsilon, \\ R &\geq \rho - \epsilon, \\ P_e' &\leq \epsilon, \\ P_e'' &\leq \epsilon. \end{aligned} \tag{12}$$

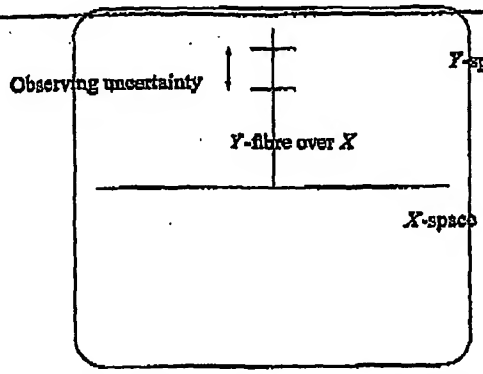


**Figure 2.** Sketch of preliminary result. The box represents the space of typical $Y$ sequences, the horizontal line the space of typical $X$ sequences, the vertical line a fibre over $X$ and the interval on the fibre the observing uncertainty. Messages are encoded by interval indices and reconstruction of the original host signal is equivalent to projection along a fibre.

The rate-distortion function $\rho_{rev}(\Delta)$ is now defined as the largest $\rho$ such that the pair $(\Delta, \rho)$ is achievable.

**Theorem 2.** *For our rate-distortion function we will show that*

$$\rho_{rev}(\Delta) = \max H(Y) - H(X) - H(Y|Z), \tag{13}$$

*where the maximum is over all test-channels $P(y|x)$ such that $\mathrm{E}_{xy}[D(x,y)] \leq \Delta$.*

In the case where the attack channel is binary symmetric with transition probability $d$ we obtain $H(Y|Z) = H(Y) + H(Z|Y) - H(Z) = H(Y) + h(d) - H(Z)$. Since in this case $H(Z) \geq H(Y)$ it follows that $H(Y|Z) \leq h(d)$. In general this inequality is strict. Therefore the preliminary result is not optimal.

## 6. PROOF

The proof breaks down into two parts. We first prove achievability, to be followed by a Fano-type argument that we cannot do better than the stated upper bound (*converse*).

### 6.1. Achievability

We first consider the non-reversible case. According to Barron[9] and Willems[12] the maximum achievable rate is given by

$$R_{nrev} = \sup I(U;Z) - I(U;X) \tag{14}$$

where the sup is over all auxiliary random variables $U$ and conditional probability functions $P(u,y|x)$ for which $\mathrm{E}[D(X,Y)] \leq d$. For a given $U$ and $P$, the associated encoding and decoding procedure uses a set of codebooks $C_m$ over $U^N$, where $m$ ranges over the set off all possible messages and where $N$ is sufficiently large. To transmit a message $m$ for a given $x_1^N$, the encoder chooses a codeword $c_{m,i}$ in $C_m$ that is typical with $x_1^N$. Finally, the encoder chooses $y_1^N$ that is jointly typical with $c_{m,i}$ and $x_1^N$. The decoder operates by finding a codeword $c$ in the union of all codebooks $C_m$ that is typical with the received sequence $z_1^N$. The label of the codebook that $c$ belongs to is an estimation of the original message. For large $N$ and appropriately chosen codebooks $C_m$ the rate of this coding scheme approximates the expression in Eq. 14 to any required precision.

Returning to the reversible case, the decoder needs $NH(X|U,Z)$ bits to reconstruct $x_1^N$. The rate $R_{rev}$ remaining for auxiliary data is therefore given by

$$\begin{aligned} R_{rev} &= R_{nrev} - H(X|Z,U) \\ &= I(U;Z) - I(U;X) - H(X|Z,U) \\ &= I(U;Z) - H(X) + H(X|U) - H(X|Z,U) \\ &= I(U;Z) + I(X;Z|U) - H(X) \\ &= I(U,X;Z) - H(X). \end{aligned}$$

Now note that

$$I(U,X;Z) \le H(Z) - H(Z|U,X,Y) = H(Z) - H(Z|Y) = I(Y;Z),$$

and that this expression is reachable by choosing $U$ equal to $Y$. Therefore

$$R_{rev} = I(Y;Z) - H(X) = (H(Y) - H(X)) - H(Y|Z)$$

is achievable.

### 6.2. Converse

In this section we outline a converse.

$$\begin{aligned} \log_2(M) &= H(W) \\ &\quad \text{(Fano's Inequality)} \\ &\le H(W) - H(W, X_1^N|\hat{W}, \hat{X}_1^N) + P_E \log_2(M) + N|\mathcal{X}|P_E + 1 \\ &\quad \text{(Independence of } X \text{ and } W) \\ &= H(W, X_1^N) - H(W, X_1^N|\hat{W}, \hat{X}_1^N) - H(X_1^N) + P_E \log_2(M) + N|\mathcal{X}|P_E + 1 \\ &\quad \text{(Conditioning decreases entropy)} \\ &\le H(W, X_1^N) - H(W, X_1^N|Z_1^N, \hat{W}, \hat{X}_1^N) - H(X_1^N) + P_E \log_2(M) + N|\mathcal{X}|P_E + 1 \\ &\quad (\hat{W} \text{ and } \hat{X}_1^N \text{ depend functionally on } Z_1^N) \\ &= I(W, X_1^N; Z_1^N) - H(X_1^N) + P_E \log_2(M) + N|\mathcal{X}|P_E + 1 \\ &\quad \text{(Refinement increases mutual entropy)} \\ &\le I(W, X_1^N, Y_1^N; Z_1^N) - H(X_1^N) + P_E \log_2(M) + N|\mathcal{X}|P_E + 1 \\ &\quad ((W, X_1^N) \to Y_1^N \to Z_1^N \text{ is a Markov chain}) \\ &\le I(Y_1^N; Z_1^N) - H(X_1^N) + P_E \log_2(M) + N|\mathcal{X}|P_E + 1 \\ &\quad \text{(See Theorem 5.2.1 in Blahut[13])} \\ &\le \sum_{n=1,N} [I(Y_n; Z_n) - H(X_n)] + P_E \log_2(M) + N|\mathcal{X}|P_E + 1 \\ &= N[I(Y;Z) - H(X) + |\mathcal{X}|P_E] + P_E \log_2(M) + 1, \end{aligned} \tag{15}$$

where $X$, $Y$ and $Z$ are random variables with

$$\Pr\{(X,Y,Z) = (x,y,z)\} = \frac{1}{N} \sum_{n=1,N} \Pr\{(X_n, Y_n) = (x,y)\} Q(z|y), \tag{16}$$

where $Q$ is the attack channel and $P_E = P_E' + P_E''$. By letting $N$ go to infinity and $P_E$ to zero the required result follows.

## 7. RECURSIVE CODES

The results of the previous two sections give upper limits to what theoretically can be achieved. In this section we describe a recursive recipe for binary sources for robust reversible watermarks that yields a performance that is better than time-sharing. We sketch an outline of the construction. We recall a high-rate non-reversible embedding construction related to the constructions proposed in van Dijk and Willems.[10]

### 7.1. Non-Reversible Embedding based on Hamming codes

To describe the code consider the (7, 4, 3) binary Hamming code. Fix a certain coset $C_w$, for some message $w = 0, 1, \cdots, 7$. Consider the vector containing the 7 host symbols $x_1, x_2, \cdots, x_7$. Denote this binary vector by $x_1^7$. Now determine the composite vector $y_1^7 \in C_w$ which is closest to $x_1^7$ in Hamming-sense.

First we determine the maximum average distortion $D_*$ of this embedding method. The Hamming code is perfect and has $d_{\min} = 3$, thus we will find a word $y_1^7 \in C_w$ at Hamming distance 1 from $x_1^7$ with probability 7/8 and a word at distance 0 with probability 1/8. Hence $D_{\text{av}} = 7/8 \times 1/7 = 1/8$. The decoder determines from the vector $y_1^7$ the coset to which it belongs, hence reliable transmission is possible with rate $R = (\log_2 8)/7 = 3/7$ bit/symbol. Thus we achieve $(D_{\text{av}}, R) = (1/8, 3/7)$. The $R/D_{\text{av}}$-ratio = 24/7, which is a factor 12/7 larger than the rate-distortion ratio by bit substitution.

We can now design a series of embedding codes, based on binary Hamming codes. For a given value $m = 2, 3, \cdots$, i.e, the number of parity check equations, the codeword length is $(2^m - 1)$. Therefore

$$R = \frac{m}{2^m - 1}$$
$$D_{\text{av}} = \frac{1}{2^m}. \tag{17}$$

Hence

$$R/D_{\text{av}} = \frac{m2^m}{2^m - 1}, \tag{18}$$

which for $m \geq 2$ is better than bit substitution.

### 7.2. Recursive Reversible Embedding

The basic ingredient of the recipe is an embedding scheme $D$ as above(no requirements on reversibility) with average distortion $D_{\text{av}}$ and rate $R$. Let $x_1^N$ be a sequence from a given memoryless source, $N$ sufficiently large. The sequence is segmented into disjoint intervals of length $K$, such that the ratio $N/K$ is sufficiently large. A message $m_1$ of size $KR$ is embedded in the first segment, resulting in a segment $y_1^K$. A priori it is not possible to reconstruct $x_1^K$ from $y_1^K$, neither is the embedding method robust. In order to achieve robustness an error correcting code is applied to the first segment. In the limit, for large segments, the required number of parity check equations is equal to $Kh(d)$, where we have assumed a symmetric attack channel with parameter $d$.

The amount of information needed to reconstruct $x_1^K$ is equal to $H(X_1^K|Y_1^K)$. The proposed method embeds this reconstruction information as well as the results of the parity check equations in the first interval, leaving room for $KR - H(X_1^K|Y_1^K) - Kh(d)$ bits of auxiliary information. Similarly as for the first interval, reconstruction information and parity check equation results for the second interval are embedded in the third interval. This process is continued recursively until the one but last segment of the sequence $x_1^N$. For this last segment of $x_1^N$, the naive method of Section 3 is used to complete the construction to a fully reversible data-hiding method. The following theorem summarizes the result.

**Theorem 3.** *Let $D$ be a data-hiding method for block length $K$ with average distortion $D_{\text{av}} = \Delta$ and rate $\rho$. View $D$ as a (not necessarily memoryless) test channel from sequences $x_1^N$ to sequences $y_1^N$. Let $C$ be the recursive construction of above. Then $C(D)$ is a reversible data-hiding scheme with average distortion $\Delta$ and rate $\rho - H(X_1^K|Y_1^K)/K - h(d)$.*

As a simple application of this theorem, we consider again the binary Bernouilli source of Section 3 with $p_0 = 0.9$ and attack channel with parametet $d = 0.05$ with $h(d) = 0.2864$. As in van Dijk and Willems [10] we consider a data-hiding method $D$ constructed from a Hamming codes. In particular we consider as a first example a Hamming code of length 3. The method given in [10] embeds auxiliary information into a sequence $x_1^N$ by modifying in each disjoint triple of symbols at most one sample. The modification is such that the embedded information can be read from the two bits in the syndrome computed over the triple. Assuming that the auxiliary bits of the auxiliary data have a random distribution, the average distortion and rate are given by $D_{\text{av}} = 1/4$ and $\rho = H(Y_1^3|X_1^3)/3 = 2/3$, respectively. The conditional entropy term $H(X_1^3|Y_1^3)$ is easily computed, resulting in a reversible data-hiding construction with distortion 1/4 and rate $0.3786 - 0.2864 = 0.0922$.

**Remark 1.** *Note that the test channels derived from the Hamming codes are memoryless when viewed as a block channel. However, the channel is not memoryless when viewed as a channel on separate symbols.*

## 8. CONCLUSION

We have proven some fundamental results on the capacity of robust reversible data-hiding schemes. We also have given a practical code construction that outperforms the classical time-sharing solution and highlights the importance of efficient non-reversible data-hiding schemes as well as the. We have applied these results to the example of a simple Bernouilli binary source. On a high abstraction level, the result of this paper can be summarized by saying that an optimal robust reversible data-hiding exploits (i) the side information available from the received data (an original sequence $x_1^N$ is reconstructed from $NH(X|Y)$ bits, not from $NH(X)$ bits) and (ii) error correcting codes to achieve resilience to attack channels.

## REFERENCES

1. T. Kalker and F. Willems, "Capacity bounds and constructions for reversible data-hiding," in *Proceedings of the International Conference on Digital Signal Processing*, 1, pp. 71–76, June 2002.
2. D. Maas, T. Kalker, and F. Willems, "A recursive code construction for reversible data-hiding," in *Proceedings of the 2002 ACM workshops on Multimedia*, Dec, 2002.
3. B. Chen, *Design and Analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems*. PhD thesis, Massachusetts Institute of Technology, June 2000.
4. B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Transaction on Information Theory* 47, pp. 1423 – 1443, May 2001.
5. P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," http://www.ifp.uiuc.edu/moulin, 1999.
6. M. Costa, "Writing on dirty paper," *IEEE Transaction on Information Theory* 47, pp. 439 – 441, May 1983.
7. S. Gelfand and M. Pinsker, "Coding for a channel with random parameters," *Problems of Control and Information Theory* 9, pp. 19–31, 1980.
8. C. Heegard and A. E. Gamal, "On the capacity of computer memory with defects," *IEEE Transactions on Information Theory* 29, pp. 731 – 739, 1983.
9. R. Barron, *Systematic Hybrid/Analog/Digital Signal Coding*. PhD thesis, Massachusetts Institute of Technology, June 2000.
10. M. van Dijk and F. Willems, "Embedding information in grayscale images," in *Proceedings of the 22nd Symposium on Information Theory in the Benelux*, pp. 147 – 154, (Enschede), May 2001.
11. J. Fridrich, M. Goljan, and R. Du, "Lossless data embedding for all image formats," in *Proceedings of SPIE, Security and Watermarking of Multimedia Contents*, (San Jose), 2000.
12. F. Willems, "An information theoretical approach to information embedding," in *Proceedings of 21st Symposium on Information Theory in the Benelux*, pp. 255 – 260, (Wassenaar, The Netherlands), May 2000.
13. R. E. Blahut, *Principles and Practice of Information Theory*, Addison Wesley, 1991.

PHNL030009EPQ 9 23.01.2003

Robust Reversible Data Hiding

**Claims:**

1. A method for robust reversible datahiding by including error correcting (ECC) data in the embedded auxiliary data stream, where the ECC data is used to achieve robustness for the recovery of the original signal and/or the embedded remaining auxiliary data.

2. A method for robust reversible datahiding as described in ID610066 where each segment is used to store error correcting data for a previous segment on top of the restoration and auxiliary message data.

3. A method of data hiding as described hereinbefore.

4. An arrangement for data hiding as described hereinbefore.

**Abstract:**

Many methods for reversible watermarking are highly fragile in the sense that the slightest modification of watermarked content prohibits the recovery of both the original signal as well as the embedded auxiliary data. In this invention we present a method for achieving robust reversible watermarks.